图书情报工作
第 64 卷 第 13 期 2020 年 7 月

阿尔茨海默病基因-疾病关联的知识挖掘*

■ 王雪[1,2] 武俊伟[3] 陈观群[4] 李燕琼[2] 马路[1]
[1] 首都医科大学医学人文学院 北京 100069 [2] 首都医科大学宣武医院图书馆 北京 100053
[3] 中国人民解放军总医院医学信息室 北京 100853 [4] 首都医科大学宣武医院神经内科 北京 100053

摘 要:[目的/意义]对阿尔茨海默病(AD)进行基因-疾病关联挖掘,以捕捉潜力研究方向。[方法/过程]基于 LBD 理论构建开放式知识发现架构,结合 MeSH 词表、DisGeNET 等医学术语、组学数据对 PubMed 中 AD 文献进行知识挖掘,采用关联规则与算法排序等方法对部分基因重合的强关联主题共现疾病和优先候选基因进行筛选,结合时间切片和其他 LBD 工具对比加以验证。[结果/结论]对 88 334 篇 AD 文献进行基因-疾病识别,并与 2 120 种 AD 基因进行匹配;以 XYZ 分析视角对识别出的 992 种主题共现疾病及 11 899 种候选基因进行关联排序;精炼 10 种强关联疾病与 25 种优选候选基因,结合文献报道加以论述。通过 LBD 挖掘目标疾病-共现疾病-基因之间潜在关联,可快速捕捉潜力研究方向,缩小基因测序范围,为新研究假设的生成提供重要指导依据。

关键词:知识发现 基因组学 阿尔茨海默病 实体识别 数据挖掘 排序算法 时间分析
分类号:G250 R745 R319.1
DOI:10.13266/j.issn.0252-3116.2020.13.016

目前,痴呆已成为老年人群致死和致残的主要疾病之一[1]。阿尔茨海默病(Alzheimer disease,AD)作为痴呆的首要病因,更是 21 世纪全球医疗卫生所面临的巨大挑战之一[2]。2015 年,全球 60 岁及以上人群的 AD 等痴呆患病率高达 5.2%,患病人数预计将在 35 年内翻倍递增[3]。AD 致残率高,患者晚期丧失独立生活能力且完全依赖于他人的持续性照护[1,4],经估算,其费用成本几乎占据全球 GDP 的 1.09% [5],给家庭与社会都带来了沉重负担。复杂的发病机制使得 40 年来该领域难以有所突破,治疗药物仍以对症为主而未能改变疾病进程[6]。因此,明确 AD 发病的危险因素并开展早期干预或预防,是延缓 AD 发病的有效途径之一[1]。

遗传因素作为除年龄外最明确的 AD 危险因素,近年来相关研究取得了一系列进展。尽管 β 淀粉样蛋白假说长期主导着诊疗方向的发展,但基于连锁分析、全基因组关联研究(Genome-wide association study,GWAS)、大规模并行重测序(Massively parallel sequencing,MPS)等技术展开的基因组学研究结果揭示了一系列促成 AD 的生物学过程,并提出新的治疗靶点[7],为探讨 AD 风险的遗传学成因、解释多因素复杂性奠定了基础。虽然这些结果在发病机制与治疗方案设计等方面的作用有限,人们仍需对 AD 新基因阐明、基因分析对疾病预防的潜在影响等遗传学研究保持乐观[8]。

已发表的科研论文中蕴含着大量生物医学知识,包括经试验(或实验)验证且被广泛接纳的“既有知识”,以及尚未被普遍关注且研究基础薄弱的“新兴知识”。虽然研究者倾向于使用既有知识体系来解释疑问,但对新兴知识的系统分析与实践验证更有利于将思维转化为可检验假设,从而激发学科内的深度挖掘与学科间的协同合作[9-10]。1986 年,D. Swanson 提出基于文献的知识发现模式(literature-based discovery,LBD)[11],尝试以自动化或半自动化方式从现有文献中发现新的、有意义的知识关联[11-13],可用于药物副作用监测、疾病新疗法研究以及候选疾病基因识别等发现过程[14]。LBD 理论应用于 AD 知识发现展现出丰富层次,包括从基因[15-16]、蛋白分子[17-19]、代谢产

* 本文系首都医科大学宣武医院院级管理课题“基于科技影响力排行的医院重点学科影响力分析”(项目编号:XWGL-2019003)和首都医科大学宣武医院院级教学课题“基于元素养理论的医学生信息素养教学路径研究”(项目编号:2019XWJXGG-10)研究成果之一。
作者简介:王雪(ORCID:0000-0002-5675-6726),助理馆员,硕士研究生;武俊伟(ORCID:0000-0002-0806-8160),信息工程师,助理工程师,硕士研究生;陈观群(ORCID:0000-0002-8133-834X),博士研究生;李燕琼(ORCID:0000-0002-1481-3593),馆长,副研究馆员,本科生;马路(ORCID:0000-0001-9147-5746),教授,博士,博士生导师,通讯作者,E-mail:malulib@ccmu.edu.cn。
收稿日期:2020-01-03 修回日期:2020-02-28 本文起止页码:120-132 本文责任编辑:王传清

王雪, 武俊伟, 陈观群, 等. 阿尔茨海默病基因－疾病关联的知识挖掘[J]. 图书情报工作,2020,64(13):120－132.

物[20－21]及疾病药物[9, 22－24]等角度入手分析,以探测AD在遗传变异、基因表型、蛋白细胞生理病生等方面的潜力研究方向。虽产生一定成果,但仍存在诸如缺乏外部验证、数据适用范围小,结果解释困难等问题[25－26]。

本研究基于LBD理论的开放式知识发现架构(见图1),通过AD文献挖掘关联疾病,结合组学数据库中基因疾病信息推测AD潜在候选基因,并采用时间切片和其他LBD工具对比加以验证,以期为后续明晰AD发病机制、扩展诊治思路提供一定参考。

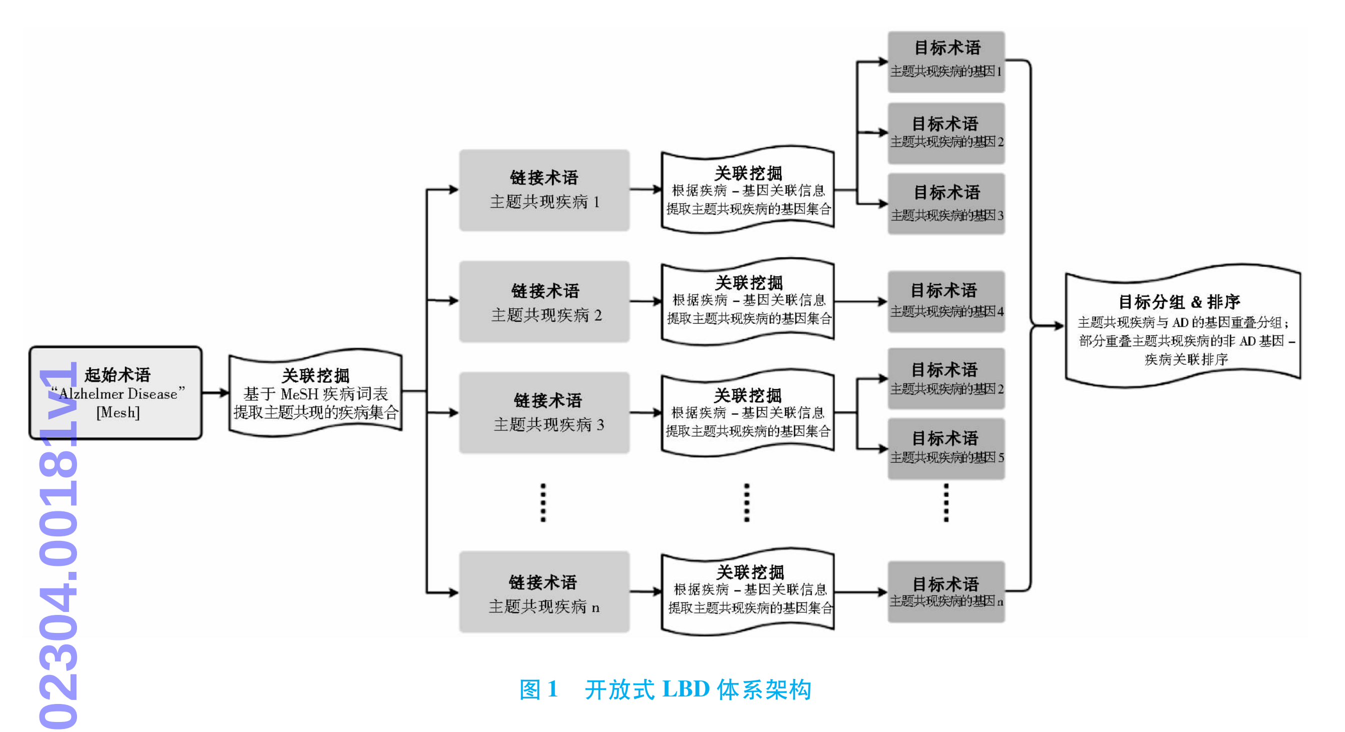


图1　开放式LBD体系架构

1　数据来源及方法

1.1　数据来源

PubMed作为当今国际上生物医学领域最权威的数据库[27],至今收录文献超过3 027万篇,其海量文献所蕴含的生物医学知识俨然构成一座浩瀚的知识宝库。医学主题词表(Medical Subject Headings, MeSH)是由美国国立医学图书馆(National Library of Medicine, NLM)编制的分层制受控词表,可精准、快速地揭示文献中生物医学概念,从而保证PubMed中海量文献的有效检索[28]。词共现(term co-occurrence)指表征文献主题的词,如关键词、标题词或主题词等共同出现在一篇文章中[29]。词共现关系是分析文献知识内容关联、挖掘知识价值的重要手段[30],常被用于预测疾病与基因之间的关联[31]。既往研究表明,对PubMed文献中MeSH进行共现分析可成功复制D. Swanson发现[32－33]。本研究基于MeSH的主题词共现(以下简称"主题共现")方式,对PubMed中AD文献的疾病主题词进行识别,为挖掘知识关联奠定基础。

AD研究的最大挑战之一是破译其发病的潜在机制。分子医学的不断发展使生物医学研究能够有效回答有关基因－疾病关联的问题[26],使用文本挖掘、多数据源集成等方式自动抓取科研文献中疾病候选基因并对其进行优先排序是获取疾病分子机制信息的策略之一[34]。当下,大量组学信息被整合在公共网络平台上,如GeneCards、UniProtKB、PharmGKB等根据基因组学、蛋白组学或药物基因组学对疾病遗传学进行注释的集成数据库;对基因组、表型和环境信息资源的综合利用,能够加深研究者对疾病机制的理解[35]。因此,应巧妙地整合此类基因疾病数据集,通过基于查询项、关联词与数据库术语的三者共现关系,结合基于规则的模式识别算法来实现基因优先排序[34, 36],从而为二代测序方向提供思路。笔者在对20种常见生物信息学数据库进行调查后,根据疾病范围、数据可获取性等指标筛选其中6种平台(见表1),汇总AD基因－疾病关联(gene-disease associations, GDAs)数据。其中,DisGeNET在识别基因和疾病词汇表方面展现出更优的全面性与灵活性,且能友好支持MeSH、UMLS、ICD9-CM等术语标识符下疾病的注释分类[35]。基于此,本研究选择从DisGeNET导出全部GDAs,作为识别上述MeSH主题共现疾病的基因注释表,为挖掘潜在知识关联提供线索。

表 1 部分生物信息学数据库/平台简介

序号	名称	类别	所属机构	数据库简介	部分特征或应用范围	相关出版物	相关链接
1	Clin Var	人类基因组变异数据库	NCBI	ClinVar 是 NCBI 主办的与疾病相关的人类基因组变异数据库。它整合了 dbSNP、dbVar、Pubmed、OMIM 等多个数据库在遗传变异和临床表型方面的数据信息,形成一个标准的、可信的遗传变异－临床相关的数据库	自主分类方式	M. Landrum 等. ClinVar at five years: Delivering on the promise.	https://www.ncbi.nlm.nih.gov/clinvar/
2	DisGeNET: Database of Gene-Disease Associations	人类疾病相关的公共可用基因和变体发现平台	GRIB	DisGeNET 是一个发现平台,包含与人类疾病相关的最大公共可用基因和变体之一。DisGeNET 集成了专家策划的存储库、GWAS 目录、动物模型和科学文献中的数据。DisGeNET 数据通过受控词表和社区驱动的本体进行统一注释。另外,提供了几个原始指标以帮助确定基因型 － 表型关系的优先次序	基因疾病关联强大	J. Piñero 等. The DisGeNET knowledge platform for disease genomics: 2019 update.	http://www.disgenet.org/
3	HPA: Human Protein Atlas	人类蛋白质图谱数据库	KAW	Human Protein Atlas 数据库提供全部 24 000 种人类蛋白质的组织和细胞分布信息,并免费提供给公众查询。用免疫组化的技术,检查每一种蛋白质在 48 种人类正常组织、20 种肿瘤组织、47 个细胞系和 12 种血液细胞内的分布和表达,其结果用至少 576 张免疫组化染色图表示,并经专业人员阅读和标引	由 3 个部分组成:组织 Altas 显示在人体内的蛋白质在所有主要的组织和器官的分布;细胞 Altas 显示蛋白质在单细胞中的亚细胞定位;病理学图谱显示蛋白质水平对癌症患者存活的影响	M. Uhlén 等. Proteomics. Tissue-based map of the human proteome.	http://www.proteinatlas.org/
4	HPO: Human Phenotype Ontology	人类表型本体数据库	Monarch Initiative	HPO 提供了人类疾病中遇到的表型异常的标准化词汇,每个术语描述了表型异常,例如房间隔缺损等。HPO 目前包含超过 13 000个术语和超过 156 000 个遗传性疾病注释,与其他项目开发了用于表型驱动的差异诊断、基因组诊断和转化研究的软件	可提供基因本体的信息下载	S. Köhler 等. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources.	https://hpo.jax.org/app/
5	MalaCards	人类疾病注释综合库	WIS	MalaCards 从 68 个数据源中提取的带注释疾病的综合纲要,它在 15 个部分中描绘了各种各样的注释主题,包括摘要、症状、解剖背景、药物、基因检测,变异和出版物。别名和分类部分反映了一种用于在经常发生冲突的来源之间整合疾病名称的算法,可提供有效的注释合并	整合了疾病的别名、基因本体等信息;整合了 GeneCards 资源	N. Rappaport 等. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search.	https://www.malacards.org/
6	OMIM: Online Mendelian Inheritance in Man	人类基因遗传数据库	NLM/JHUSM	OMIM 是人类孟德尔遗传数据库(线上版)(online Mendelian Inheritance in Man)的简称。这是一个持续更新的关于人类基因和遗传紊乱的数据库,主要着眼于遗传性的基因疾病,包括文本信息和相关参考信息、序列纪录、图谱和相关其他数据库	主要关注人类基因变异和表型性状之间的关系	J. Amberger 等. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes.	https://omim.org/

注:NCBI:National Center for Biotechnology Information,美国国家生物技术信息中心;GRIB:Research Programme on Biomedical Informatics,西班牙生物医学信息学研究计划项目组;KAW:Knut and Alice Wallenberg Foundation,瑞典 Knut and Alice Wallenberg 基金会;WIS:Weizmann Institute of Science,以色列魏茨曼科学研究学院;NLM:National Library of Medicine,美国国立医学图书馆;JHUSM :Johns Hopkins University School of Medicine,美国美国约翰霍普金斯医学院

1.2 研究方法

具体步骤见数据处理流程图(见图 2),包括:①通过 PubMed 检索“Alzheimer Disease”[Mesh]下载 AD 主题相关文献,随后进行去重整理。②从 NLM 的 FTP 站点(http://www.nlm.nih.gov/mesh/meshhome.html)获取 MeSH 术语词表,提取 Diseases Category 所在的 C 类疾病词表信息(Tree Numbers:C)。③将 MeSH 疾病词表与上述 AD 文献主题词进行匹配,识别与 AD 共现的疾病信息。④根据 DisGeNET 提取的全病种 GDAs 术语词表对主题共现疾病进行基因－疾病关联,获取疾病的全部基因。⑤根据 6 个平台汇总后的 AD 基因集合对主题共现疾病的全部基因进行识别,获取与 AD 基因重合、包含 AD 基因或无 AD 基因的疾病列表。运行过程中,使用 VBA 编程实现识别、匹配等数据处理,所处理数据存储在 Access 数据库中。

王雪，武俊伟，陈观群，等.阿尔茨海默病基因－疾病关联的知识挖掘[J].图书情报工作,2020,64(13):120－132.

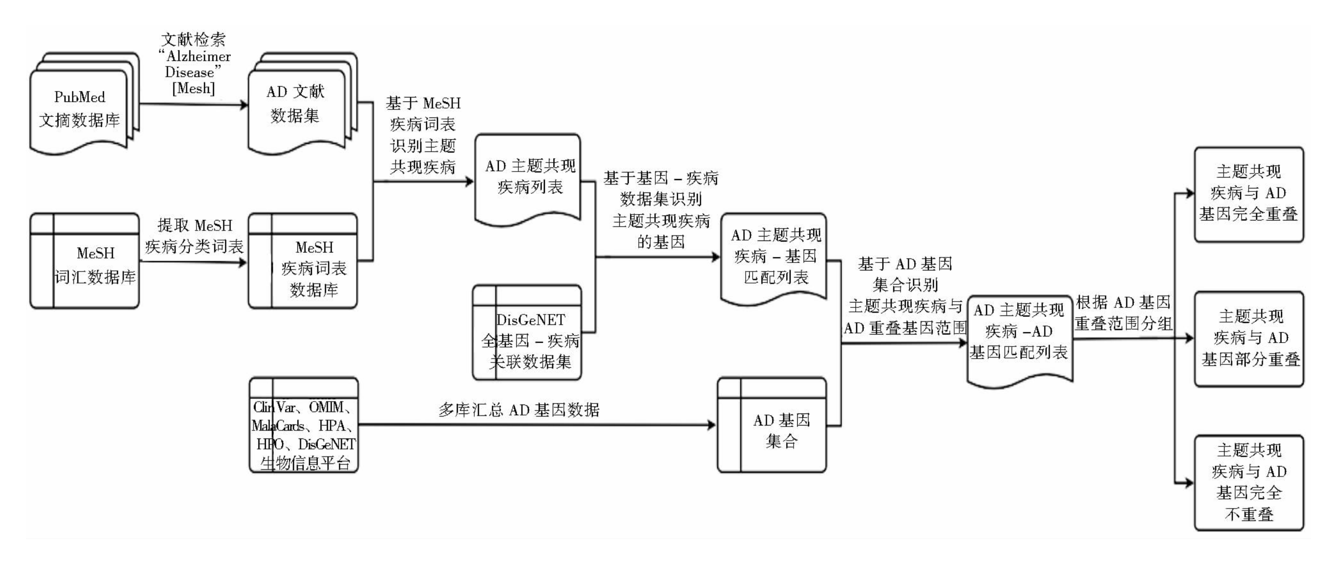


图 2　数据处理流程

2　结果与分析

2.1　整体结果

检索日期为2019年7月31日，共检索到88 334篇1945年以来AD主题相关文献，利用MeSH C类词表(下载日期为2019年6月23日)中11 648类/4 818种疾病对其进行疾病实体识别，匹配出166 946次/1 639种AD主题共现疾病。根据628 685条DisGeNET全病种GDAs(下载日期为2019年8月2日)匹配出其中1 125种，获取关联基因151 710条/13 891种。从Clin Var、MalaCards等6个数据库汇总AD基因共计2 120种，与主题共现疾病的关联基因进行匹配，区分每种疾病是否与AD存在相同基因：①88种疾病的135种关联基因完全与AD基因重合；②992种疾病的13 891种关联基因部分包含AD基因，涉及AD基因1 992种、非AD基因11 899种；③45种疾病的87种关联基因未包含AD基因。

2.2　相关分析

LBD理论提示只有新颖的链接才有意义。在修剪已知概念配对后，对剩余配对(即潜在发现)进行排序，以便研究人员优先探索最具潜力的研究方向[37]。因此，本研究将重点讨论与AD部分基因重合的主题共现疾病及其所涉非AD基因，通过对起始项X(AD)－链接项Y(部分基因重合的疾病)－目标项Z(疾病的其他基因)进行关联规则[38]及算法排序[39]，以实现潜在候选基因的排序，进而预测后续潜力研究方向。

2.2.1　主题共现疾病分析

以X→Y(confidence, support)关联规则[38]对AD与主题共现疾病联系进行赋值计算，通过公式(1)和公式(2)分别截取XY、YZ关联降序下前10名进行分析(见表2、表3)：

$$confidence = \frac{Dx \cap Dy}{Dx} \qquad 公式(1)$$

$$support = Dx \cap Dy \qquad 公式(2)$$

其中，Dx为AD总文献数；Dy为主题共现疾病文献数；Dx∩Dy为AD文献中主题共现疾病文献量。

表 2　AD 部分基因重合的主题共现疾病——XY 关联降序下 Top10 结果

序号	主题共现疾病英文名称	主题共现疾病中文名称	所涉文献数(篇)	XY 关联			YZ 关联				
				Rank	Support	Confidence(%)	Rank	相同基因数	不同基因数	总基因数	相同基因占比(%)
1	Dementia	痴呆	9 341	1	9 341	10.82	5	293	142	435	67.36
2	Plaque, Amyloid	淀粉样斑块	3 501	2	3 501	4.06	2	204	30	234	87.18
3	Parkinson Disease	帕金森病	3 405	3	3 405	3.94	9	530	533	1063	49.86
4	Memory Disorders	记忆障碍	2 702	4	2 702	3.13	3	32	11	43	74.42
5	Dementia, Vascular	血管性痴呆	2 401	5	2 401	2.78	1	71	8	79	89.87
6	Nerve Degeneration	神经变性	1 710	6	1 710	1.98	6	90	54	144	62.50
7	Inflammation	炎症	1 538	7	1 538	1.78	8	229	199	428	53.50
8	Lewy Body Disease	Lewy 体病	1 292	8	1 292	1.50	4	91	41	132	68.94
9	Down Syndrome	唐氏综合征	1 276	9	1 276	1.48	7	255	219	474	53.80
10	Cerebrovascular Disorders	脑血管障碍	776	10	776	0.90	10	80	97	177	45.20

表 3 AD 部分基因重合的主题共现疾病——YZ 关联降序下 Top10 结果

序号	主题共现疾病英文名称	主题共现疾病中文名称	所涉文献数(篇)	XY 关联			YZ 关联				
				Rank	Support	Confidence (%)	Rank	相同基因数	不同基因数	总基因数	相同基因占比(%)
1	Cerebral Amyloid Angiopathy	脑淀粉样血管病	618	3	618	0.72	1	43	1	44	97.73
2	Amyloid Neuropathies, Familial	家族性淀粉样神经病	11	7	11	0.01	2	11	1	12	91.67
3	Dementia, Vascular	血管性痴呆	2 401	2	2 401	2.78	3	71	8	79	89.87
4	Retinal Drusen	视网膜小疣	8	8	8	0.01	4	8	1	9	88.89
5	Toxoplasmosis, Cerebral	脑弓形虫病	1	10	1	0.00	4	8	1	9	88.89
6	Hypoxia-Ischemia, Brain	脑缺氧缺血	22	6	22	0.03	6	7	1	8	87.50
7	Neuroleptic Malignant Syndrome	安定药恶性综合征	7	9	7	0.01	6	7	1	8	87.50
8	Spinal Cord Injuries	脊髓损伤	33	5	33	0.04	6	7	1	8	87.50
9	Plaque, Amyloid	淀粉样斑块	3 501	1	3 501	4.06	9	204	30	234	87.18
10	Tauopathies	Tau 病变	463	4	463	0.54	10	111	18	129	86.05

淀粉样斑块(Plaque, Amyloid)与血管性痴呆(Dementia, Vascular)在两种排序方式下均占据前列。从文献数量(X→Y)来讲,淀粉样斑块(4.06%,3 501)一直以来都是 AD 领域的主要分支;从相同基因(Y→Z)来讲,它的 AD 相同基因占比(87.18%,204)也位居前列。作为 AD 标志性神经病理学改变,β-淀粉样蛋白(amyloid beta,Aβ)及淀粉样沉积物在 AD 发病机制中扮演着至关重要的角色。多年来,Aβ 假说认为,当细胞外聚集形成 β 淀粉样蛋白沉积物时,会触发神经退行性过程,从而导致记忆力与认知能力的丧失,继而引发 AD[40]。然而,阐明 Aβ 形成过程中有毒物质是如何产生以及该物质如何引起细胞功能障碍死亡等,仍然是个挑战。随着冷冻电子断层成像术(cryoelectron tomography,cryo-ET)等技术的提升,研究者将对 Aβ 蛋白结构、斑块聚集机制及与 AD 联系进行更深入的研究,为寻求诊断、研发延缓甚至阻碍疾病进程的药物提供新思路[40-41]。血管性痴呆(Vascular Dementia,VaD)(2.78%,2 401)是仅次于 AD 的痴呆分型,占总体的 5% 至 10%[42],由脑血管及相关病变所致脑组织血流灌注障碍,引起局部脑组织细胞损害,最终表现为认知功能障碍甚至痴呆[43]。许多老年期痴呆患者常伴 VaD 与 AD 两种病理表现,而两者共有的危险因素、Aβ 沉积现象以及一氧化氮依赖下线粒体异常活动、细胞分裂等病理因素,揭示了两者在发病机制上的共性[43-44],提示脑血管病变与神经退行性病理过程可能相互作用[45]。此外,记忆障碍、脑淀粉样血管病、Tau 病变等均列前位,提示其从研究热度、基因组学相关性上均有较高的研究价值。

2.2.2 潜在候选基因分析

考虑到 X→Z 可能会存在一个以上的中间 Y,且 X 可通过不同 Y 到达 Z,因此对 Z 进行排序时借鉴了启发式排名函数[39]的方程,见公式(3),以筛选强关联信息。利用截至 2019 年完整数据提取 AD 候选基因,结合 PubMed、Entrez 数据库掌握 AD 与潜力基因的相关研究(见表 4)。

$$\mathrm{Rank}(Z_k) = \sum_{i=1}^{m} (Sxy_i \times Sy_i\, Z_k) \qquad 公式(3)$$

其中,Z_k 为秩和后候选基因 Z 排序;S_{xy} 与 S_{yz} 为 $X \to Y_i$ 与 $Y_i \to Z_k$ 的 support 值;m 为中间概念 Y_i 数量。

截至 2019 年数据共提取 11 899 种候选基因,根据算法筛选出 25 种潜力候选基因并进行部分文献验证,结果表明:①SPP1 作为三种排序及两种关联值角度下均位列第一的基因,其关联表现异常突出。SPP1 基因是分泌磷酸蛋白 1/骨桥蛋白(Secreted Phosphoprotein 1, SPP1)的编码基因,在脑及其他多种组织中表达,参与炎症和抗凋亡过程,起细胞粘附分子和细胞因子的作用[46]。2015 年,M. Shi 等发现 SPP1 蛋白为首的脑脊液 5 肽组合标记物在区分帕金森病(Parkinson's Disease,PD)与 AD 方面具有显著特异性与敏感性[47]。随后,脑脊液、尿 SPP1 蛋白作为候选诊断标志物被用于 MCI 及 AD 前期的进展监测[48-49]。A. Rentsendorj 阐明 SPP1 蛋白可调节巨噬细胞介导下促进 Aβ 清除的过程,提出在 AD 模型中脑骨桥蛋白增加与 Aβ 减少相关[50]。W. Kamphuis 与 Z. Yin 对 APP/PS1 小鼠 CD11c+小胶质细胞、MHC II+斑块相关小胶质细胞的转录谱研究显示,SPP1 作为上调基因参与了细胞分化、系统发育等过程[51-52]。2019 年,C. Frigerio 在基因调节小胶质细胞对 Aβ 斑块的最新研究中发现,App^{NL-GF} 小鼠 Aβ 斑块的存在促进了稳态小胶质细胞向活化小胶质细胞的重新分布,SPP1 作为参与组织修复的基因会进一步区分活化小胶质细胞亚群,有利于揭

表4　2019年潜在AD候选基因信息－3种排序

序号	候选基因名称	与AD文献数量(篇)	YZ关联 Rank	YZ关联 关联值(次)	XYZ关联 Rank	XYZ关联 秩和值	序号	候选基因名称	与AD文献数量(篇)	YZ关联 Rank	YZ关联 关联值(次)	XYZ关联 Rank	XYZ关联 秩和值
1	SPP1	8	1	151	1	22 996	14	FUS	2	338	35	3	20 044
2	TAC1	2	23	80	2	20 825	15	C19orf12	1	1 437	16	6	16 876
3	GPX1	10	98	55	4	19 429	16	DCTN1	0	947	21	7	16 704
4	THBS1	8	12	87	5	18 849	17	JPH3	0	757	24	8	16 674
5	MECP2	13	34	74	10	16 511	18	DNAH8	0	164	46	9	16 520
6	CXCL10	15	5	110	11	16 399	19	KRAS	4	2	118	572	3 690
7	MIR21	1	7	102	16	15 522	20	FOXP3	18	3	112	210	5 896
8	SELE	4	27	78	18	15 207	21	BRAF	4	4	111	839	3 169
9	CCR2	42	25	79	30	13 811	22	CTLA4	2	6	106	810	3 373
10	NF1	2	93	56	37	12 838	23	IL2RA	1	8	101	283	5 015
11	PTHLH	0	66	62	39	12 768	24	HLA-C	7	9	96	182	6 312
12	NCAM1	5	21	81	88	10 484	25	F3	1	10	93	376	4 435
13	TXN	3	47	68	98	10 333							

注：No 1－13代表XYZ秩和与YZ关联值降序均在前100名的基因信息；No14－18代表除上述部分基因外XYZ秩和前10的基因信息；No19－25代表除上述部分基因外YZ关联值前10的基因信息

示AD小胶质细胞的病理特征[53]。自2007以来，SPP1作为基因调控产物－蛋白分子生物标记物方面的研究报道陆续产出，但直接探讨基因转录、表达及参与AD发病机制的研究仍然薄弱，结合既往研究成果与本研究数据判断，SPP1与AD相关研究值得继续深入。②其中PTHLH作为不同关联值下均位居前列的候选基因，却与AD研究暂无交集，而相似情况也发生在其他候选基因上(见表5)，提示虽无文献支撑，但该类基因与其他神经退行性病变或神经系统疾病(Nervous System Disease，NSD)仍有报道，可通过与AD强关联NSD进一步挖掘候选基因的潜力研究方向。

表5　2019年AD特殊候选基因及PubMed所涉基因与AD/NSD研究文献量－XYZ关联降序

序号	特殊候选基因名称	YZ关联 Rank	YZ关联 关联值(次)	XYZ关联 Rank	XYZ关联 秩和值	是否15年后加入候选范围	与AD研究数量(篇)	与NSD研究数量(篇)	相关检索式
1	DCTN1	947	21	7	16 704	否	0	75	(DCTN1) AND "Nervous System Diseases"[Mesh]
2	JPH3	757	24	8	16 674	否	0	26	(JPH3) AND "Nervous System Diseases"[Mesh]
3	DNAH8	164	46	9	16 520	否	0	1	PubMed Links for Gene (Select 1769) AND "Nervous System Diseases"[Mesh]
4	PTHLH	66	62	39	12 768	否	0	20	(PTHLH) AND "Nervous System Diseases"[Mesh]
5	TRNS2	702	25	74	10 889	是	0	2	(MT-TS2 OR TRNS2 OR MTTS2 OR TRNS-2) AND "Nervous System Diseases"[Mesh]
6	TRNW	948	21	81	10 711	是	0	14	(TRNW OR MTTW OR MT-TW) AND "Nervous System Diseases"[Mesh]
7	ATP6V1A	3 283	8	130	9 819	是	0	5	(ATP6V1A OR HO68 OR VA68 OR VPP2 OR Vma1 OR ARCL2D OR ATP6A1 OR IECEE3 OR ATP6V1A1) AND "Nervous System Diseases"[Mesh]
8	RNF216	2 659	10	135	9 559	是	0	12	(RNF216 OR RNF-216 OR CAHH OR U7I1 OR TRIAD3 OR UBCE7IP1) AND "Nervous System Diseases"[Mesh]

注：特殊候选基因，指PubMed上该基因与AD无直接相关文献；NSD，Nervous System Diseases，神经系统疾病；2019年候选基因列表中抽取4种特殊候选基因，并罗列2015年后新划入候选范围的4种特殊候选基因进行分析

3　评估

LBD的评估具有挑战性[54]，所捕获的新发现尚未在任何领域发布，难以验证其有效性[55]。然而，了解发现结果的可靠性至关重要，主要可通过黄金标准集和评估指标来进行衡量[54]。研究者常使用基线对比、

经典发现复制、时间切片、专家/用户评估或实验/试验验证等技术评估其结果，并结合信息检索等定量指标检验其性能[12, 55]。

3.1 时间切片

3.1.1 评估方案

时间切片是 LBD 的主要评估方法[12]，根据截止日期将数据集分为发现前、发现后两段，通过训练前段数据以生成发现，再将后段数据用作测试集以开发黄金标准集来评估发现[12, 56-57]。其中，黄金标准的制定互不相同，主要取决于关联术语的评估方式[54]，形式也不拘泥于后段数据提取，如专家意见、专利试验等均可创建为黄金标准集[12]。本研究选取生物信息数据库中新确认的 AD 基因－疾病关联为黄金标准集，相较语义提取及关系判定更加精准。

医学领域的发现需要时间，合理的截止日期对假设转变为事实至关重要，但日期的划分暂无标准且具有高度主观性[58]。根据 AD 文献发展趋势，本研究以增长趋于平稳的 2014 至 2015 年作为时间分割范围，划定 2014 年 12 月 31 日为时间截点提取截点前期 AD 候选基因，结合截点后期新确认 AD 基因集合进行验证。同时，采用精确度(Precision，P 值)、召回度(Recall，R 值)、F 度量(F-Measure，F 值)等信息检索指标，分别对全数据及前 20 准确值区间做定量评估。

3.1.2 整体评估结果

截至 2014 年 12 月 31 日，相关文献数据共提取候选基因 10 564 种。经与 2015 年后 AD 基因集合对比，其中 380 种预测成功，整体 R 值 = 0.825 7，P 值 = 0.035 9(见表 6)。借助 11 点插值 P-R 值曲线[54]观察各梯度 R 值与 P 值变化趋势，调整后插值平均精度(Average interpolated precision，AiP)上升至 0.1260，提示排序对整体性能确有影响。图 3 显示，R 值在 0.002 6到 0.100 3 区间 P 值下降幅度明显，提示列表中 R 值在 10% 以内的预测结果(预测 247，成功 38)精准性较强。以加权 F 值曲线[59-60]综合权衡 P-R 值，图 3 中 F1/F2 曲线显示，当均衡或优先考虑预测较全，P 值在 R = 0.401 1 时表现最佳(预测 1 404，成功 152)；若优先考虑预测较准，F0.5 曲线显示 P 值在 R = 0.200 5时表现最佳(预测 533，成功 76)，提示当更关注预测结果精确率时，所浏览预测基因数无需过多。

3.1.3 区间评估结果

考虑到多数科研人员不会浏览所有发现，因此评估前 k 个位置中关联比例很重要[55]。全数据 P-R 曲线(见图 3)揭示发现结果排序靠前被成功预测的几率更大。截取集合中成功预测的前 20 个候选基因作为检测区间再次评估，检索 PubMed 以补充预测失败基因与 AD 强关联的支撑文献，以对时间切片的部分排序进行校正(见表 7)。

表 6 2015 年前潜在 AD 候选基因的预测评估

	全部预测集合	前 20 种正确值集合	
		经时间切片验证	经文献证据校正
预测基因数	10 564	119	48
预测正确基因数	379	20	20
精确度(P)	0.035 876 562	0.168 067 227	0.416 666 667
召回度(R)	0.825 708 061	—	—
插值平均精度(AiP)	0.125 980 887	—	—
平均精度(AP)	—	0.253 750 617	0.693 277 398

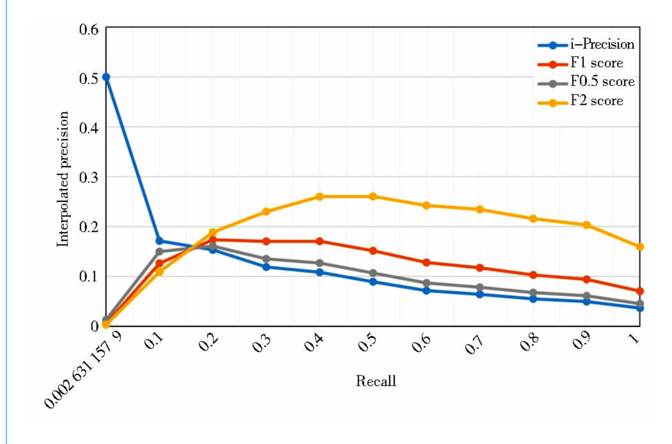


图 3 2015 年前 AD 候选基因全数据的 11 点插值精确度－召回度曲线

经时间切片验证，浏览截至 119 种预测基因时可获取 20 个成功值，平均精确度(Average Precision，AP)为 0.253 8，较全数据集有所提升。图 4 显示，R 值 = 0.3(预测 21，成功 6)，达到平衡点[60](Break-Even Point，BEP)，提示在浏览前 21 个候选基因时所收获成功基因的概率最高。F 值曲线显示，若更加关注预测精准率，浏览至第 13 种预测基因时即可获取最佳精确度(预测 13，成功 5)；两者均权衡时，则需要浏览更多(预测 55，成功 12)。

文献检索发现，2015 年以前部分预测失败的 AD 关联基因研究已经发表，(如 ATP5PD，PMID：23857120)，因其关联信息未在 2015 年前被 DisGeNET 等数据库收录，故在时间切片测试中未能识别出该类基因信息，提示除生物信息数据库所收录的信息外，PubMed 文献中仍有散落的关联基因信息。介于此，在忽略文献结论权威性的情况下，本研究严格划定文献纳入标准，筛选 AD 患者/动物模型/体外相关基因调控表达且揭示其正/逆向关联的研究作为证据资源，以

表 7　2015 年前 AD 候选基因预测正确 Top20 结果－XYZ 关联降序

序号	候选基因名称	预测 Rank	YZ 关联值(次)	XYZ 关联秩和值	经时间切片验证		经文献证据校正	
					确认 Rank	确认年份	确认 Rank	部分关键文献
1	SPP1	1	133	17 274			1	PMID:31018141
2	EGF	2	101	16 678	1	2016	2	
3	TAC1	3	71	15 647			3	PMID:26402107
4	FMR1	4	37	14 305	2	2015	4	
5	ATXN2	7	33	13 341	3	2016	5	
6	GPX1	9	45	12 805			6	PMID:29246792
7	CHMP2B	10	16	12 305	4	2015	7	
8	CXCL10	12	97	11 980			8	PMID:30529693
9	GDF15	13	46	11 968	5	2016	9	
10	PRDX5	15	18	11 689			10	PMID:28358580
11	PARK7	16	15	11 680			11	PMID:30889441
12	NAGLU	18	13	11 422			12	PMID: 20040070
13	VPS13A	20	7	10 918			13	PMID: 26825611
14	MEF2C	21	17	10 874	6	2016	14	
15	MIR21	22	73	10 765			15	PMID:29635890
16	CALCA	27	65	10 125	7	2018	16	
17	DPP4	32	57	9 631	8	2016	17	
18	PLP1	39	22	9 073			18	PMID:29110684
19	TYROBP	47	14	8 582	9	2016	19	
20	RBM8A	48	7	8 570			20	PMID:31816601

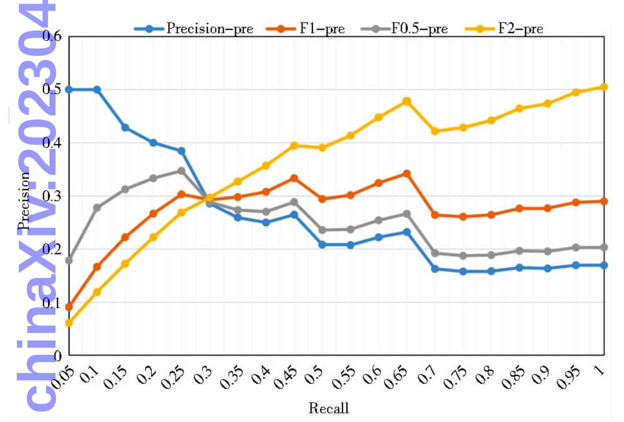


图 4　2015 年前 AD 候选基因预测正确 Top20 的精确度－召回度曲线(时间切片)

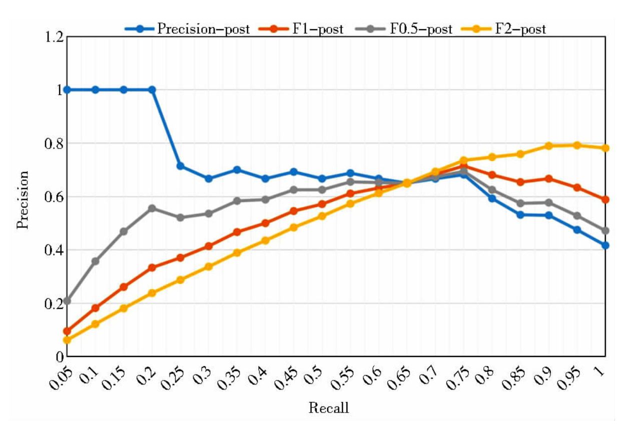


图 5　2015 年前 AD 候选基因预测正确 Top20 的精确度－召回度曲线(文献校正)

校正排序结果(见图 5)。经校正,在浏览至第 48 种基因时即可获取前 20 个成功值(见表 7),AP 值＝0.693 3,较校正前有大幅提升,揭示了文献查阅对于预测结果解读的必要性。经调整,当 R 值＝0.65 时(预测 20,成功 13)达到平衡点,与校正前平衡点所需预测数目相近。F 值曲线说明,在不优先考虑召回率的情况下,浏览 22 个预测基因即可获得较高准确率。

3.2　其他评估方式

其他 LBD 模型往往难以回溯历史数据。本研究采用即时输出结合文献预估的方式,与同样基于 XYZ 理论及关联规则的 BITOLA[39] 进行横向对比。双方以 AD 为 X 起始概念,截取 XY 关联值降序下前 50 种疾病为 Y 中间概念,罗列 Z 候选基因列表。BITOLA 和本研究方案分别筛选出 4 211 种、5 252 种候选基因。表 8 显示当截至前 20 个预测结果时,两者的排序大相径庭,而在对方列表中多位居其后。BITOLA 预测结果中有密切支撑文献(提示基因调控表达或生物标记物)的候选基因数低于本研究方法,而其预测的 TIMP1、EPO 等基因在 DisGeNET 中已被标为 AD 关联基因。目前的已有数据难以完整展现两者性能差异,需继续扩展区间测定范围,尝试运行多种主题或利用其他黄金标准集来完善评估。

表 8 2019 年 AD 候选基因预测 Top20 结果对比 – XYZ 关联降序

本研究方案							BITOLA 结果							
预测排序	基因名称	XY 关联值	XYZ 秩和值	对方排序	是否有相关文献	文献估量（篇）	预测排序	基因名称	XY 关联值	XYZ 秩和值	对方排序	是否有相关文献	文献估量（篇）	补充说明
1	SPP1	18	20 687	1 823	是	9	1	IL5	14	12 740 670	1 758	是	1	/
2	FUS	13	19 535	/	是	7	2	MYCN	5	10 790 840	1 655	否	0	/
3	TAC1	14	19 525	2 536	是	2	3	PARP9	10	10 710 710	/	否	0	19 全数据集有预测 – 5656
4	GPX1	11	18 244	243	是	11	4	ARL4C	9	5 769 360	/	否	0	19 全数据集有预测 – 6651
5	THBS1	11	17 279	/	是	8	5	MOG	14	5 155 030	4 278	是	5	/
6	C19orf12	5	16 384	/	是	1	6	TIMP1	20	4 941 350	/	是	8	DisGeNET – 2003
7	DCTN1	7	16 290	1 890	否	0	7	EPX	18	4 914 640	/	否	0	/
8	JPH3	9	16 080	328	否	0	8	EPO	18	4 594 120	/	是	30	DisGeNET – 2012
9	PRDX5	6	16 010	1 166	是	1	9	ERBB2	7	4 353 730	/	是	5	DisGeNET – 2018
10	DNAH8	9	15 375	/	否	0	10	RAPGEF5	15	3 926 370	306	否	0	/
11	ND5	8	14 989	/	是	1	11	TBC1D9	15	3 926 370	192	否	0	/
12	MECP2	11	14 957	460	是	13	12	ANG	15	3 899 660	/	是	3	DisGeNET – 2016
13	FTL	5	14 889	1 407	是	3	13	CD86	15	3 739 400	/	是	7	DisGeNET – 2016
14	CYTB	7	14 844	/	是	1	14	AGTR2	17	3 525 720	/	是	7	DisGeNET – 2009
15	CXCL10	12	14 841	/	是	15	15	SUB1	15	3 472 300	5 033	否	0	/
16	ATP13A2	4	14 745	/	否	0	16	ADIPOQ	11	3 472 300	506	是	14	/
17	VPS13C	5	14 741	/	是	1	17	KDR	16	3 418 880	524	是	1	/
18	PARK7	7	14 630	/	是	14	18	MDM2	10	3 418 880	/	是	5	DisGeNET – 2014
19	LY6E	5	14 620	/	否	0	19	AFP	14	3 392 170	910	否	0	/
20	NAGLU	5	14 309	154	是	2	20	CD80	13	3 312 040	/	是	3	DisGeNET – 2016

数据或文本挖掘的研究成果也部分佐证了本研究的预测。F. Yao 通过分析 AD 脑组织基因表达数据结合 iTRAQ 实验，发现 AD 患者尿液 SPP1 蛋白差异表达，将其报道为早期 AD 的尿液生物标记物[49]。Y. Cruz-Rivera 等利用微阵列数据集结合旅行商问题(Traveling Salesman Problem，TSP)路径对 AD 患者与对照组神经元的差异表达进行分析，发现 FTL 处在最相关循环中，可作为潜在 AD 生物标记物[61]。不同研究所得的相近结论[15,49,61–62]将为结果预测的有效性提供一定依据。

4 讨论

本研究以 LBD 理论入手，对 PubMed 中 AD 文献进行基因–疾病关联挖掘。运用 LBD 框架进行诠释[31]：本研究主要以 AD、主题共现疾病及关联基因为对象(Objects)，利用 MeSH 主题共现、疾病实体识别及基因匹配等方式构建链接(Links)，通过多种组学数据库中 GDAs(Additional sources)挖掘 AD 与主题共现疾病的差异基因集合以推理疾病下基因之间的隐藏关联(Inference)，并结合关联规则与排序算法(Intermediary)筛选强关联主题共现疾病及 AD 优先候选基因，进而为 AD 新研究的假设提供线索。

早期的知识发现研究主要集中在数据库领域，随着新兴技术与应用模式的涌现，其研究重点开始转向对非结构化数据(文本数据)的知识提取[30]。LBD 作为知识发现重要分支之一，在生物信息挖掘领域的研究愈发广泛，其技术也在不断精进[55]，相关进展包括：①数据类型：将 LBD 应用于专利、病例报告等论文之外的类型；②分析单元：使用 UMLS、MeSH、Entrez Gene 等受控词汇提取概念以促成跨学科的知识发现；③处理流程：在手动检测基础上提出如基于概念、关系、图或链接预测等自动处理技术；④过滤机制：进行词级过

滤前对文章、段落及语句消除噪声关联以缩小发现范围;⑤排序技术:除常规统计排序外,使用机器学习模型对潜在关联排序;⑥结果输出:在展示关联排序列表基础上采用基于语义类型、图形可视化、矩阵可视化或发现途径等技术;⑦发现评估:采用定量与定性结合的多重评估方式。LBD 研究已在新药研发、药物再利用和药物不良事件预测中得以实践[12],但始终缺乏模仿概念之间联系真正形成的能力,需通过整合逻辑和优化推理机制进一步完善 LBD 认知过程,以便更好地理解复杂关联[63]。此外,LBD 发现是基于既有文本形成的探索性假设,始终需要最终用户决定接受与否[31],建立专家评估或开展用户交互研究将为验证其有效性提供可靠依据[12, 55, 63]。

本研究在既往 LBD 研究基础上,尝试整合不同来源组学数据以更好地满足关联发现需要,并将其作为系统评估标准以确保黄金标准集的准确性。结合 GDAs 数据对 AD 主题共现疾病及所涉候选基因进行双向分析,设置多重排序提炼强关联疾病和优先候选基因,更利于精准把握潜力基因的预测范围,从而有效指导科研方向,节省时间与成本。时间切片、文献校正及 LBD 系统横向对比的多重评估结果表明,用户在浏览本研究预测的前 20－22 个 AD 候选基因即可达到较优精确率,在一定程度上体现了性能和效能。

受到数据库及词表范围限制,本研究仅对 Medline 文献中被主题标引的疾病进行了实体识别,若能将文献范围扩展至 PubMed 全库或其他数据库,同时应用 Emtree、UMLS 等更多类型医学术语建立映射加强识别,可增强候选基因的基因－共现疾病关联性。本研究的识别规则主要基于主题词共现,分析结果很难提供有力的证据来解释发病机制的因果关系[64],但并不影响对于疾病与基因关联的提取[65]。本研究以基因为着手点,未能对其他生物医学概念(蛋白、细胞、代谢产物等)进行探讨,未来将考虑使用更多关系类型实体强化发现联系,构建异构网络,结合本体或可视化图谱等技术进一步延伸 AD 知识发现研究。此外,考虑到结果外部验证、实际数据适用性等 LBD 共同问题,未来可与临床、基础科研团队合作继续深化相关研究。

5 结语

生物信息学的快速发展为神经科学做出了重要贡献,将基因型与表型联系起来用于新关联的发现仍是 AD 等神经退行性疾病病因学研究的主要挑战之一[9, 64]。本研究期望通过对 AD 进行知识挖掘以快速捕捉更具潜力的研究方向,进一步缩小基因测序范围、辅助科研人员聚焦更有价值的研究目标,从而为新研究假设的诞生提供重要指导建议,为后续明晰 AD 发病机制、扩展诊治思路提供重要参考依据。

参考文献:

[1] 中国痴呆与认知障碍诊治指南写作组, 中国医师协会神经内科医师分会认知障碍疾病专业委员会. 2018 中国痴呆与认知障碍诊治指南(七):阿尔茨海默病的危险因素及其干预[J]. 中华医学杂志, 2018, 98(19): 1461－1466.

[2] SCHELTENS P, BLENNOW K, BRETELER M M, et al. Alzheimer's disease[J]. Lancet, 2016, 388(10043): 505－517.

[3] PRINCE M J, WIMO A, GUERCHET M M, et al. World Alzheimer report 2015 － the global impact of dementia[M]. London: Alzheimer's Disease International, 2015.

[4] TAYLOR C A, GREENLUND S F, MCGUIRE L C, et al. Deaths from Alzheimer's disease - United States, 1999－2014[J]. MMWR-morbidity and mortality weekly report, 2017, 66(20): 521－526.

[5] JIA J, WEI C, CHEN S, et al. The cost of Alzheimer's disease in China and re-estimation of costs worldwide[J]. Alzheimers & dementia, 2018, 14(4): 483－491.

[6] PATTERSON C. World Alzheimer report 2018 － the state of the art of dementia research: new frontiers[M]. London: Alzheimer's Disease International, 2018.

[7] VERHEIJEN J, SLEEGERS K. Understanding Alzheimer disease at the interface between genetics and transcriptomics[J]. Trends in genetics, 2018, 34(6): 434－447.

[8] VAN CAUWENBERGHE C, VAN BROECKHOVEN C, SLEEGERS K. The genetic landscape of Alzheimer disease: clinical implications and perspectives[J]. Genetics in medicine, 2016, 18(5): 421－430.

[9] MALHOTRA A, YOUNESI E, GURULINGAPPA H, et al. 'Hypothesisfinder': a strategy for the detection of speculative statements in scientific text[J]. Plos computational biology, 2013, 9(7): e1003117.

[10] HENRY S. Indirect relatedness evaluation and visualization for literature based discovery[D]. Virginia: Virginia Commonwealth University, 2019.

[11] SWANSON D R. Fish oil, raynaud's syndrome, and undiscovered public knowledge[J]. Perspectives in biology and medicine, 1986, 30(1): 7－18.

[12] HENRY S, MCINNES B T. Literature based discovery: models, methods, and trends[J]. Journal of biomedical informatics, 2017, 74:20－32.

[13] COHEN T, SCHVANEVELDT R W. The trajectory of scientific discovery: concept co-occurrence and converging semantic distance[J]. Studies in health technology and informatics, 2010, 160(1): 661－665.

[14] HRISTOVSKI D, RINDFLESCH T, PETERLIN B. Using literature-based discovery to identify novel therapeutic approaches [J]. Cardiovascular & hematological agents in medicinal chemistry, 2013, 11(1): 14-24.

[15] KIM Y H, BEAK S H, CHARIDIMOU A, et al. Discovering new genes in the pathways of common sporadic neurodegenerative diseases: a bioinformatics approach [J]. Journal of Alzheimers disease, 2016, 51(1): 293-312.

[16] KAWALIA S B, RASCHKA T, NAZ M, et al. Analytical strategy to prioritize Alzheimer's disease candidate genes in gene regulatory networks using public expression data [J]. Journal of Alzheimers disease, 2017, 59(4): 1237-1254.

[17] GUBIANI D, FABBRETTI E, CESTNIK B, et al. Outlier based literature exploration for cross-domain linking of Alzheimer's disease and gut microbiota [J]. Expert systems with applications, 2017, 85:386-396.

[18] GRECO I, DAY N, RIDDOCH-CONTRERAS J, et al. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation [J]. Journal of translational medicine, 2012, 10:217.

[19] MALHOTRA A, YOUNESI E, BAGEWADI S, et al. Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease [J]. Genome medicine, 2014, 6(11): 97.

[20] SMALHEISER N R, SWANSON D R. Linking estrogen to Alzheimer's disease: an informatics approach [J]. Neurology, 1996, 47(3): 809-810.

[21] YETISGEN-YILDIZ M, PRATT W. Using statistical and knowledge-based approaches for literature-based discovery [J]. Journal of biomedical informatics, 2006, 39(6): 600-611.

[22] SMALHEISER N R, SWANSON D R. Indomethacin and Alzheimer's disease [J]. Neurology, 1996, 46(2): 583.

[23] LI J, ZHU X Y, CHEN J Y. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts [J]. Plos computational biology, 2009, 5(7): e1000450.

[24] CHEN R, LIN H F, YANG Z H. Passage retrieval based hidden knowledge discovery from biomedical literature [J]. Expert systems with applications, 2011, 38(8): 9958-9964.

[25] ZHANG R, SIMON G, YU F. Advancing Alzheimer's research: a review of big data promises [J]. International journal of medical informatics, 2017, 106:48-56.

[26] RAJA K, PATRICK M, GAO Y, et al. A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries [J]. International journal of genomics, 2017, 2017:6213474.

[27] 刘群, 孙昌朋, 王谦, 等. 入选 PubMed 数据库对提升医学期刊国际影响力的作用 [J]. 中国科技期刊研究, 2015, 26(12): 1344-1347.

[28] 刘菊红, 于建荣, 缪有刚. 基于 MeSH 词表和共词分析的疾病本体半自动构建方法研究 [J]. 现代情报, 2009, 29(3): 208-211.

[29] 张云秋, 冷伏海. 非相关文献知识发现的理论基础研究 [J]. 中国图书馆学报, 2009, 35(4): 25-30.

[30] 阮光册. 主题模型与文本知识发现应用研究 [M]. 上海:华东师范大学出版社, 2018.

[31] SEHGAL A, QIU X, SRINIVASAN P. Analyzing LBD methods using a general tramewerk [C]//BRUZA P, WEEBER M. Literature-based discovery. Berlin: Springer, 2008:75-100.

[32] STEGMANN J, GROHMANN G. Hypothesis generation guided by co-word clustering [J]. Scientometrics, 2003, 56(1): 111-135.

[33] STEGMANN J, GROHMANN G. Advanced information retrieval for hypothesis generation [C] // Society for Information Science. International workshop on webometrics, informetrics and scientometrics. Roorkee: Central Library, Indian Institute of Technology, 2004: 334-346.

[34] ONO T, KUHARA S. A novel method for gathering and prioritizing disease candidate genes based on construction of a set of disease-related MeSH (R) terms [J]. BMC bioinformatics, 2014, 15:179.

[35] PINERO J, QUERALT-ROSINACH N, BRAVO A, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes [J]. Database-the journal of biological databases and curation, 2015:bav028.

[36] RAPPAPORT N, FISHILEVICH S, NUDEL R, et al. Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect [J]. Biomedical engineering online, 2017, 16(s1): 72.

[37] SHUI Q Y. Big data analysis for bioinformatics and biomedical discoveries [M]. Portland: CRC Press, 2016.

[38] FAYYA D, USAMA M. Advances in knowledge discovery and data mining [M]. California: AAAI Press, 1996.

[39] HRISTOVSKI D, PETERLIN B, MITCHELL J A, et al. Using literature-based discovery to identify disease candidate genes [J]. International journal of medical informatics, 2005, 74(2/4): 289-298.

[40] MAKIN S. The amyloid hypothesis on trial [J]. Nature, 2018, 559(7715): s4-s7.

[41] IADANZA M G, JACKSON M P, HEWITT E W, et al. A new era for understanding amyloid structures and disease [J]. Nature reviews molecular cell biology, 2018, 19(12): 755-773.

[42] Alzheimer's Association. Vascular dementia [EB/OL]. [2019-12-24]. https://www.alz.org/alzheimers-dementia/what-is-dementia/types-of-dementia/vascular-dementia.

[43] 吴佳慧. 阿尔茨海默病和血管性痴呆的病理机制及相关临床研究比较 [J]. 浙江医学, 2019, 41(11):1227-1231.

[44] ASHRAF G M, CHIBBER S, MOHAMMAD, et al. Recent updates on the association between Alzheimer's disease and vascular

dementia [J]. Medicinal chemistry, 2016, 12(3): 226－237.
[45] 中国医师协会神经内科分会认知障碍专业委员会,《中国血管性认知障碍诊治指南》编写组. 2019 年中国血管性认知障碍诊治指南[J]. 中华医学杂志, 2019, 99(35):2737－2744.
[46] WUNG J K, PERRY G, KOWALSKI A, et al. Increased expression of the remodeling and tumorigenic associated factor osteopontin in pyramidal neurons of the Alzheimer's disease brain [J]. Current alzheimer research, 2007, 4(1): 67－72.
[47] SHI M, MOVIUS J, DATOR R, et al. Cerebrospinal fluid peptides as potential Parkinson disease biomarkers: a staged pipeline for discovery and validation [J]. Molecular & cellular proteomics, 2015, 14(3): 544－555.
[48] BEGCEVIC I, BRINC D, BROWN M, et al. Brain-related proteins as potential CSF biomarkers of Alzheimer's disease: a targeted mass spectrometry approach [J]. Journal of proteomics, 2018, 182:12－20.
[49] YAO F, HONG X, LI S, et al. Urine-based biomarkers for Alzheimer's disease identified through coupling computational and experimental methods [J]. Journal of Alzheimers disease, 2018, 65(2): 421－431.
[50] RENTSENDORJ A, SHEYN J, FUCHS D T, et al. A novel role for osteopontin in macrophage-mediated amyloid-β clearance in Alzheimer's models [J]. Brain behavior and immunity, 2018, 67:163－180.
[51] KAMPHUIS W, KOOIJMAN L, SCHETTERS S, et al. Transcriptional profiling of CD11c-positive microglia accumulating around amyloid plaques in a mouse model for Alzheimer's disease [J]. Biochimica et biophysica acta-molecular basis of disease, 2016, 1862(10): 1847－1860.
[52] YIN Z, RAJ D, SAIEPOUR N, et al. Immune hyperreactivity of Aβ plaque-associated microglia in Alzheimer's disease [J]. Neurobiology of aging, 2017, 55:115－122.
[53] SALA FRIGERIO C, WOLFS L, FATTORELLI N, et al. The major risk factors for Alzheimer's disease: age, sex, and genes modulate the microglia response to Aβ plaques [J]. Cell reports, 2019, 27(4): 1293－306. e1－e6.
[54] YETISGEN-YILDIZ M, PRATT W. Evaluation of literature-based discovery systems [C]//BRUZA P, WEEBER M. Literature-based discovery. Berlin: Springer, 2008: 101－13.
[55] THILAKARATNE M, FALKNER K, ATAPATTU T. A systematic review on literature-based discovery workflow [J]. PeerJ computer science, 2019, 5:e235.
[56] HRISTOVSKI D, STARE J, PETERLIN B, et al. Supporting discovery in medicine by association rule mining in Medline and UMLS [J]. Studies in health technology and informatics, 2001, 84(2): 1344－1348.
[57] HENRY S, MCINNES B T. Indirect association and ranking hypotheses for literature based discovery [J]. BMC bioinformatics, 2019, 20(1): 425.
[58] YETISGEN-YILDIZ M, PRATT W. A new evaluation methodology for literature-based discovery systems [J]. Journal of biomedical informatics, 2009, 42(4): 633－643.
[59] HRIPCSAK G, ROTHSCHILD A S. Agreement, the f-measure, and reliability in information retrieval [J]. Journal of the American Medical Informatics Association, 2005, 12(3): 296－298.
[60] CARTERETTE B, VOORHEES E M. Overview of information retrieval evaluation [C]//LUPU M, MAYER K, TAIT J, et al. Current challenges in patent information retrieval. Berlin: Springer, 2011: 69－85.
[61] CRUZ-RIVERA Y E, PEREZ-MORALES J, SANTIAGO Y M, et al. A selection of important genes and their correlated behavior in Alzheimer's disease [J]. Journal of Alzheimers disease, 2018, 65(1): 193－205.
[62] CIFUENTES R A, MURILLO-ROJAS J. Alzheimer's disease and HLA-A2: linking neurodegenerative to immune processes through an in silico approach [J]. Biomed research international, 2014: 791238.
[63] GOPALAKRISHNAN V, JHA K, JIN W, et al. A survey on literature based discovery approaches in biomedical domain [J]. Journal of biomedical informatics, 2019, 93:103141.
[64] HOFMANN-APITIUS M, BALL G, GEBEL S, et al. Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders [J]. International journal of molecular sciences, 2015, 16(12): 29179－29206.
[65] PLETSCHER-FRANKILD S, PALLEJA A, TSAFOU K, et al. DISEASES: text mining and data integration of disease-gene associations [J]. Methods, 2015, 74:83－89.

作者贡献说明:

王雪:起草研究框架,制定研究方法,收集资料,整理数据,撰写与修订论文;

武俊伟:参与制定研究方法,编程与数据处理,参与论文修订;

陈观群:参与制定研究方法,提供学术咨询,参与论文修订;

李燕琼:协助论文框架制定,参与论文修订;

马路:提出论文研究思路,参与论文修订。

Knowledge Mining of Alzheimer's Disease Gene-Disease Associations

Wang Xue[1,2] Wu Junwei[3] Chen Guanqun[4] Li Yanqiong[2] Ma Lu[1]
[1] Medical Humanities School, Capital Medical University, Beijing 100069
[2] Department of Library, Xuanwu Hospital, Capital Medical University, Beijing 100053
[3] Medical Information Section, Chinese PLA General Hospital, Beijing 100853
[4] Department of Neurology, Xuanwu Hospital, Capital Medical University, Beijing 100053

Abstract: [**Purpose/significance**] To explore the gene-disease association of Alzheimer's disease (AD) in order to capture the potential research directions. [**Method/process**] An open knowledge discovery framework was constructed based on LBD theory. Combined with MeSH thesaurus, DisGeNET and other medical terms and group data, knowledge mining was carried out in AD literatures in PubMed. Association rules and algorithm sorting were used to screen strongly associated MeSH terms co-occurrence diseases and priority candidate genes for partial gene coincidence, results of time slicing and comparison with other LBD tools were used to verify them. [**Result/conclusion**] 88 334 AD literatures were identified and matched with 2 120 AD genes, 11 899 candidate genes and 992 comorbidity genes were identified according to XYZ analysis, 10 strongly associated co-occurrence diseases and 25 preferred candidate genes were refined and discussed in combination with literature reports. Mining the potential associations between target disease, co-occurrence diseases and genes by LBD can quickly capture the potential research directions, narrow the scopes of gene sequencing, and provide important guidance for the generations of new research hypotheses.

Keywords: literature based discovery genomics Alzheimer's disease entity recognition data mining sorting algorithm time analysis